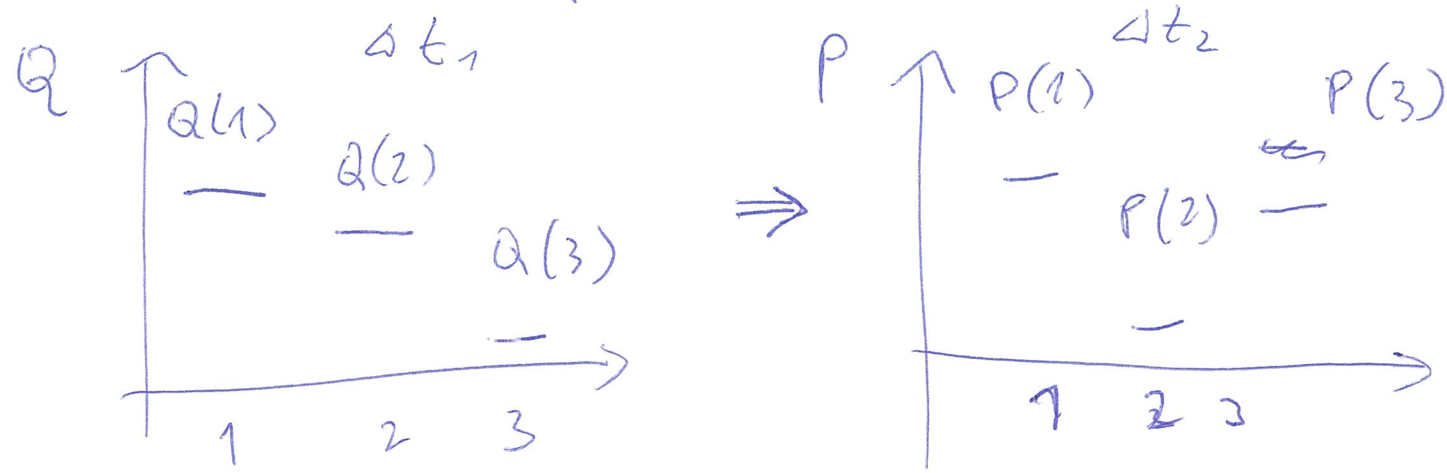
-1- KL - Divergence (Kullback-Leibler)



$$\text{proportion} = \frac{P(x)}{Q(x)}$$

$P(x)$ → new p.d.f.

$Q(x)$ → reference p.d.f.

$KL(P \| Q)$ → how P 'has changed' relative to Q

has changed ≡ ~~≠~~

≡ different

$$1: \frac{P(1)}{Q(1)} = 1$$

$$2: \frac{P(2)}{Q(2)} = \frac{1}{4}$$

$$3: \frac{P(3)}{Q(3)} = 4$$

How to use it?

regression function

$$\text{Average (metric)}: \frac{1}{l}\sum_l \frac{P(l)}{Q(l)} =$$

$$= \frac{1 + \frac{1}{4} + 4}{3} = 1.75$$

scalar metric

- Sometimes, asymmetry is a very desired property
- Simple average is not,
- Mean value is not ok also because it is easily biased with outliers → salaries are not analysed with mean (median)
- How to balance this situation

e.g., $\frac{1}{10}$ → balancing 10

p_1 p_2

$$\frac{\sum_i p_i}{n} \rightarrow \boxed{p_1 = -p_2}$$

- if proportion is 1 → $p_3 = 1$, we do not want it!

$$\frac{P(X)}{Q(X)} \rightarrow \log \frac{P(X)}{Q(X)}$$

$$\frac{1}{i} \sum_i \log \frac{P(i)}{Q(i)}$$ → 'ok-ish' but still symmetrical

Weighting procedure → this is going to stress the change even more ≡ new values are more important in a sense (the opposite could also be true)

$$\underbrace{\sum_l P(l) \log \frac{P(l)}{Q(l)}}$$

$w_l = P(l)$

difference with respect to Q

$$KL \Rightarrow KL(P \| Q) =$$

$$\boxed{= \sum_l P(l) \log \frac{P(l)}{Q(l)}}$$

$$KL(P \| Q) = \int_{-\infty}^{+\infty} P(x) \log \frac{P(x)}{Q(x)} dx$$

[KL is always positive or 0 if $P(X) \equiv Q(X)$]

Jensen's inequality

$$E[f(X)] \geq f(E[X])$$

$$\left(\sum_{l \in L} P(l) \log \frac{P(l)}{Q(l)}\right) \geq \log\left(\sum_{l \in L} P(l) \frac{P(l)}{Q(l)}\right)$$

$$\log\left(\sum_{l\in L}\frac{P^2(l)}{Q(l)}\right) =$$

$$= -\log\left(\sum_{l\in L}\frac{P^2(l)}{Q(l)}\right)^{-1} \geq$$

$$\geq -\log\left(\frac{1}{|L|}\sum_{l\in L}\frac{P^2(l)}{Q(l)}\right) =$$

$$= \log(|L|) - \log\left(\sum_{l\in L}\frac{P^2(l)}{Q(l)}\right)$$

$$\vdots$$

$$\text{Ex. } KL(P\|Q) \leq \log(|L|)$$

Jensen - Shannon - div.

Symmetric and smoothed version of KL divergence

$P(x)$ and $Q(x)$

$$\begin{cases} JS(P\|Q) = \frac{1}{2}\left\{KL(P\|M) + KL(Q\|M)\right\} \\ M = \frac{1}{2}(P+Q) \end{cases}$$

→ average of both distributions

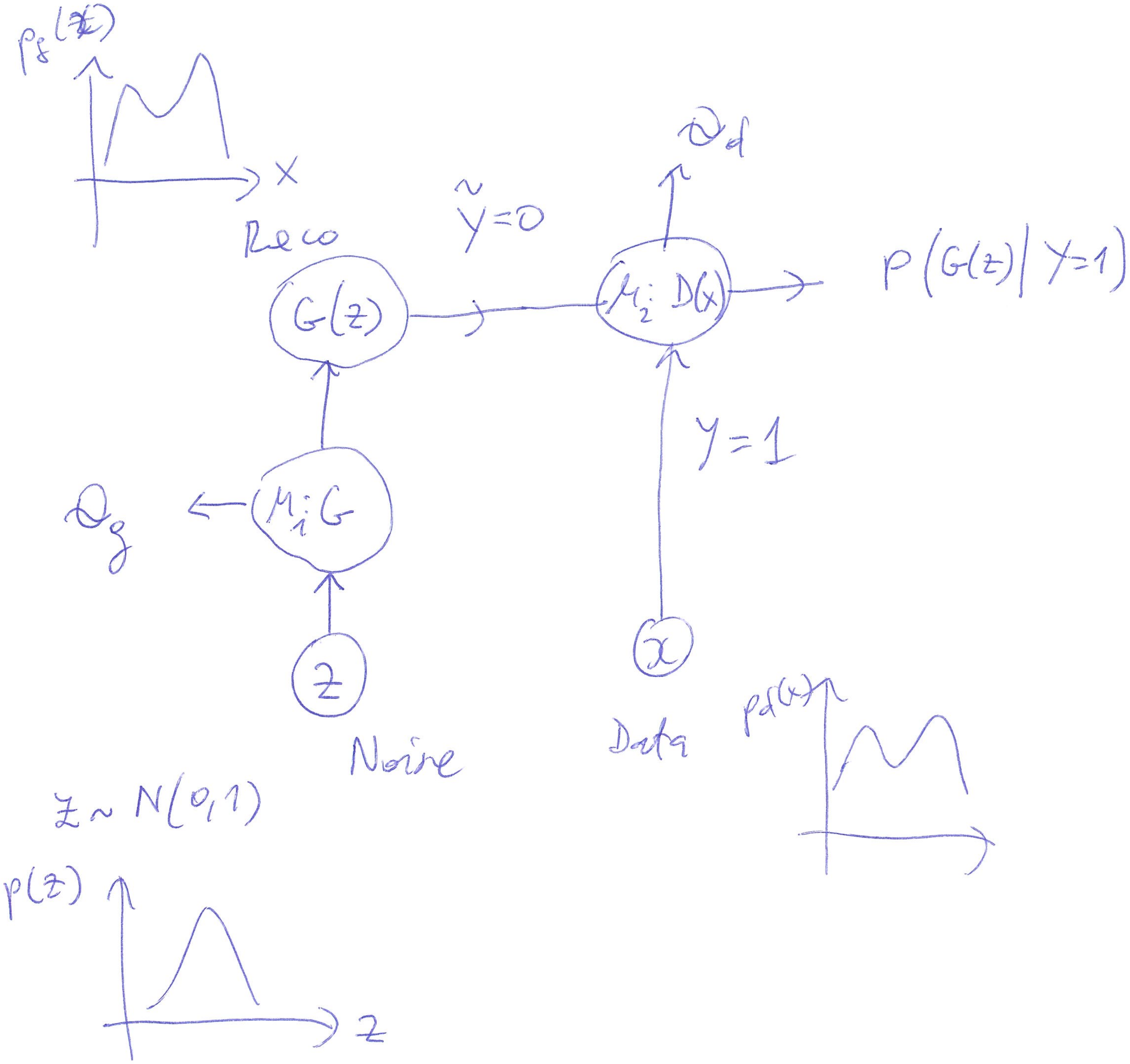
Similarity measure : small P and Q are similar , large they are different

$\max JS(P\|Q) = \log 2$

∘) useful for clustering and

∘) GAN quality improvements

-6- Why GAN works at all?



$G(z) \rightarrow p_d(x)$ Universal app. theorem (1989/1991)

G | D two mini-max players game

Target function

$$\min_{G(z)} \max_{D(x)} V(G,D) = E_{x \sim pd}\left[\log(D(x))\right] + E_{z \sim pz}\left[\log(1-D(G(z)))\right]$$

Formula similar to binary x-entropy

$\rightarrow \mathcal{L} = - y \log \tilde{y} + (1-y)\log(1-\tilde{y})$

True data $y=1$, $\tilde{y} = D(x)$

$\rightarrow \mathcal{L}_{y=1} = \log[D(x)]$

Fake data $y=0$, $\tilde{y} = D(G(z))$

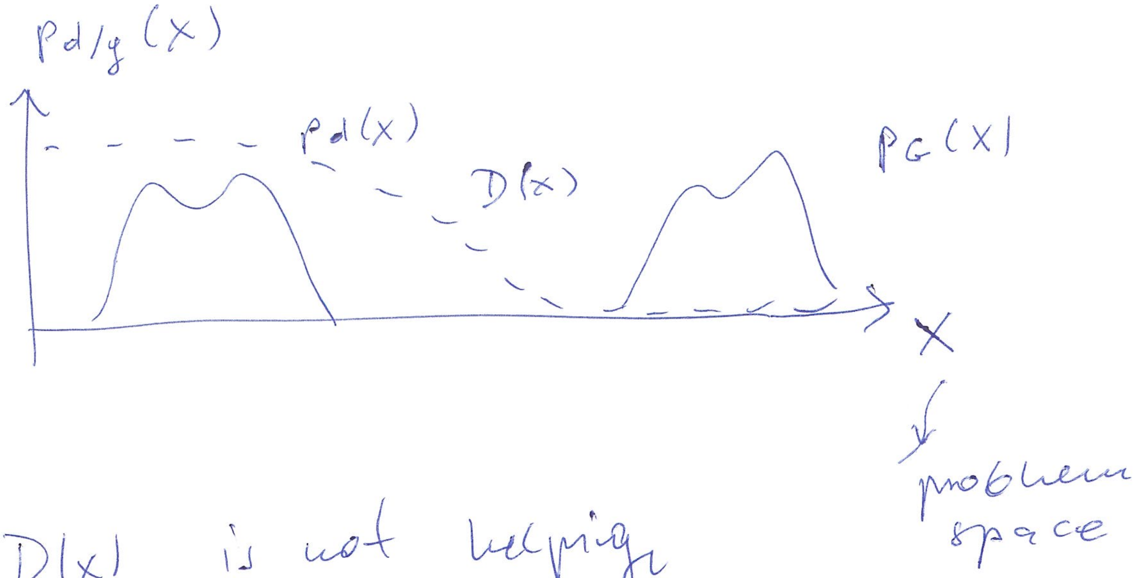
$\rightarrow \mathcal{L}_{y=0} = \log[1-D(G(z))]$

$$E[\log(D(x))] = \sum_i pd(x_i) \log(D(x))$$

$$\int pd(x) \log(D(x))\, dx$$

$$E[\log(1-D(G(z)))] = \sum_i p_z(z) \log\{1-D(G(z))\}$$

GAN's training is typically hard



D(x) is not helping in training.

- modify D gradients
- modify pd and pG distributions
- WGAN

Typical divergencies do not take into account distance between distributions

pd - pG

$$W(p_d, p_G) = \inf_{\gamma} E_{(x,y)\sim\gamma(x,y)}[\|x-y\|]$$

$\gamma \rightarrow$ that minimize the distance

$\gamma_x(x) = p_d(x)$

$\gamma_y(x) = p_G(x)$

Main loop

→ fix updates for G model

m times → inner loop updating D

·) get l instances of $y=1$ and l instances of $y=0$ data

·) update α_d params using grad ascent algo

$$\partial_{\alpha_d} \frac{1}{l} \left\{ \log(D(x)) + \log(1 - D(G(z))) \right\}$$

→ fix updates for D model

·) take l $y=0$ samples and update α_g using grad. descent

$$\partial_{\alpha_g} \frac{1}{l} \left\{ \log\left[1 - D(G(z))\right] \right\}$$

Has this a chance to converge and what the convergence mean?

Information entropy is max if all answers are equally likely

If we go away from the equal prob. the entropy goes down (we introduced predictability)

Entropy goes down we need to ask fewer questions to gues the outcome.

Amount of information in an event
$\sim$ entropy

Information theory $\rightarrow$ data compression, source coding

Surprise $\rightarrow$ probability of event

$\begin{cases} \text{high prob.} \rightarrow \text{low information} \\ \text{low prob.} \rightarrow \text{high information} \end{cases}$

$\begin{cases} \text{h.p.} \rightarrow \text{l.i.} \rightarrow \text{l.s.} \\ \text{l.p.} \rightarrow \text{h.i.} \rightarrow \text{h.s.} \end{cases}$

$\boxed{I(x)} = -\log_2(p(x))$, $x \rightarrow$ an event

unit of information in $\boxed{\text{bits}}$

if $p(x) = 1$, $I(x) = 0$

Information for R.V. X with prob. distr. $p(X)$: $H(X)$

$H(X) \rightarrow$ information entropy

$\downarrow$

The average number of bits required to represent an event drawn from the prob. distribution

$$\boxed{H(X) = -\sum p \cdot \log_2(p)} =$$

$$= \sum p \cdot \log_2\left(\frac{1}{p}\right)$$

Largest entropy $\rightarrow$ all events are equally likely

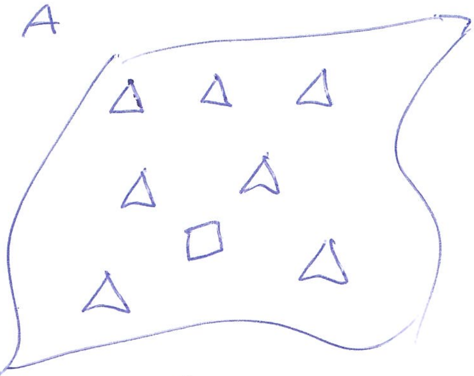
Information $\equiv$ drop in entropy

$$H(X) = -\sum_i p(x_i) \cdot \log_2(p(x_i))$$

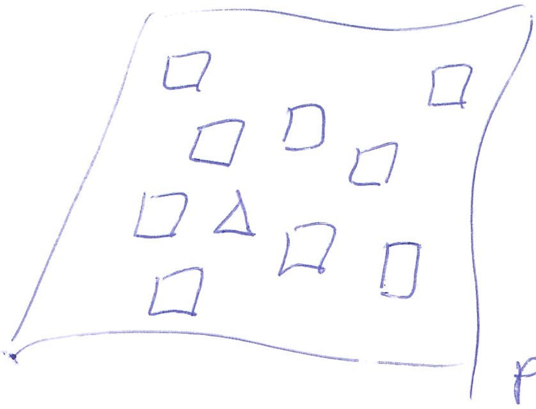
$$H(p,q) = -\sum_i p(x_i) \cdot \log(q(x_i))$$

$H(p,q) \equiv$ MIN if $p(x_i) = q(x_i)$

A



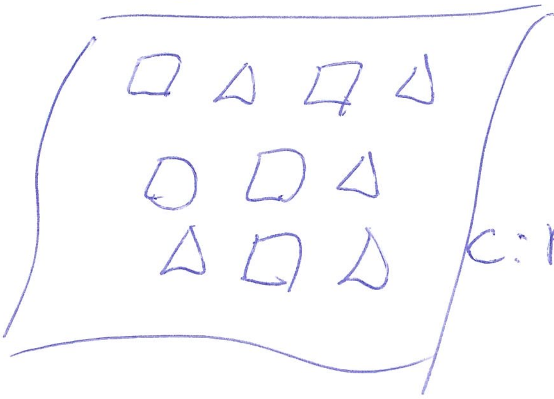
B



A: $P(\triangle)\uparrow$ $P(\square)\downarrow$

B: $P(\square)\uparrow$ $P(\triangle)\downarrow$

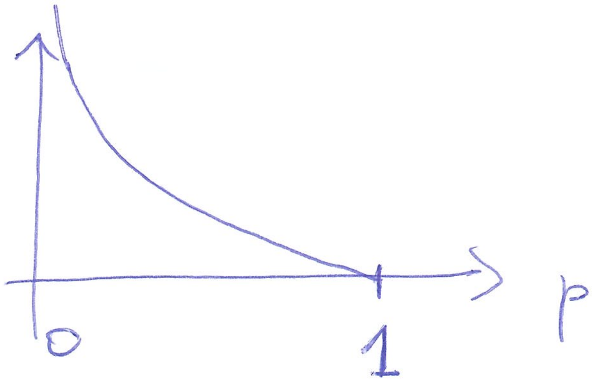
C

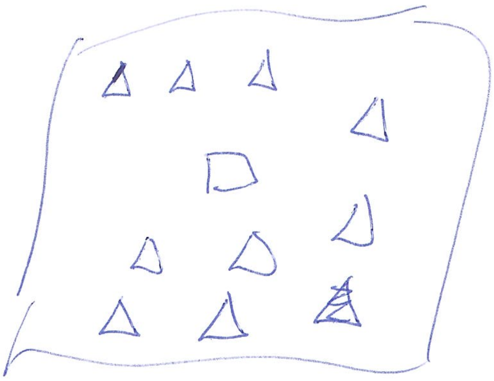


C: $P(\triangle) = P(\square)$

A: $\triangle$ h.p., l.s.
$\square$ l.p, h.s

$$\boxed{S \sim \frac{1}{p}} \qquad S = \log\left(\frac{1}{p}\right)$$





A

$$\begin{cases} P(\triangle) = \dfrac{\#\triangle}{\#All} = \dfrac{9}{10} \\ P(\square) = \dfrac{\#\square}{\#All} = \dfrac{1}{10} \end{cases}$$

Say there is no $\square$, $P(\square) = 0$

$P(\triangle) = 1$, $S = \dfrac{1}{p} = \dfrac{1}{1} = \underline{\underline{1}}$

Not good, we would like to get $\boxed{\text{surprise} \equiv 0}$

$$S = \log_2\left(\frac{1}{p}\right) = -\log_2(p)$$

$P(\triangle) \rightarrow$ getting $\triangle$ is no surprise

$P(\square) \rightarrow$ never happens, never eletive in this case

$p(\triangle) = 0.8, \quad p(\square) = 0.1$

$S_\triangle = \log_2(1) - \log_2(0.8) = 0.15$

$S_\square = \log_2(1) - \log_2(0.1) = 3.32$

$S_\square > S_\triangle$

Any sequence:

$S_1 = \triangle, \triangle, \square, \triangle \quad , \quad p(\triangle, \triangle, \square, \triangle) = 0.8 \times 0.8 \times 0.1 \times 0.8$

$S_1 = -\log_2(p(S_1)) = -[3 \cdot \log_2(0.8) + \log_2(0.1)] =$

$= 3 \cdot 0.15 + 3.32 = 3.81$ [bits]

This works for any sequence

For instance 100 draws with return

$\underbrace{[0.8 \times 100]}_{\text{prob. of } \triangle \text{ in 100 draws} \approx \text{ expected number of } \triangle} \times \underbrace{0.15}_{\text{surprise of obs. } \triangle} + \underbrace{[0.1 \times 100]}_{\text{expected \# of } \square \text{ in 100 draws}} \times 3.32 = 46.7$

$\frac{1}{100}$ ≡ entropy

expected value of surprise/information is entropy

$$H(A) = 0.47$$

For the fair coin

$$H(f.c.) = \Big\{\underbrace{[0.5 \times 100]}_{\#\,\text{heads}} \times 1 + \underbrace{[0.5 \times 100]}_{\#\,\text{tails}} \times 1\Big\} \frac{1}{100}$$

$= 1$ entropy of a fair coin

$$H(x) = \sum_i x_i \cdot p(X = x_i)$$

x_i: surprise or information

$p(X = x_i)$: prob. of observing specific surprise

$$H(x) = \sum_i p(x_i) \cdot \log\left(\frac{1}{p(x_i)}\right) =$$

$$= -\sum_i p(x_i) \cdot \log(p(x_i))$$

$$S = k_b \cdot \ln \Omega$$

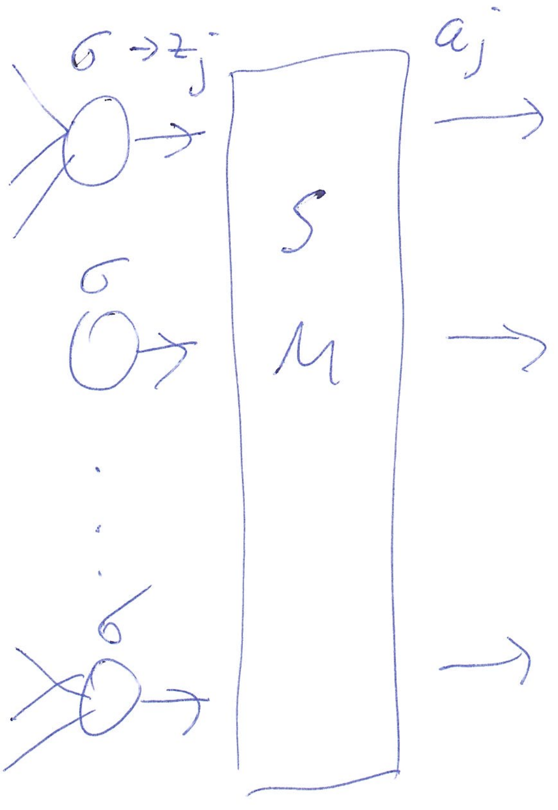
disorder

$$H(x) = -\sum_i p(x_i) \log_2 (p(x_i))$$

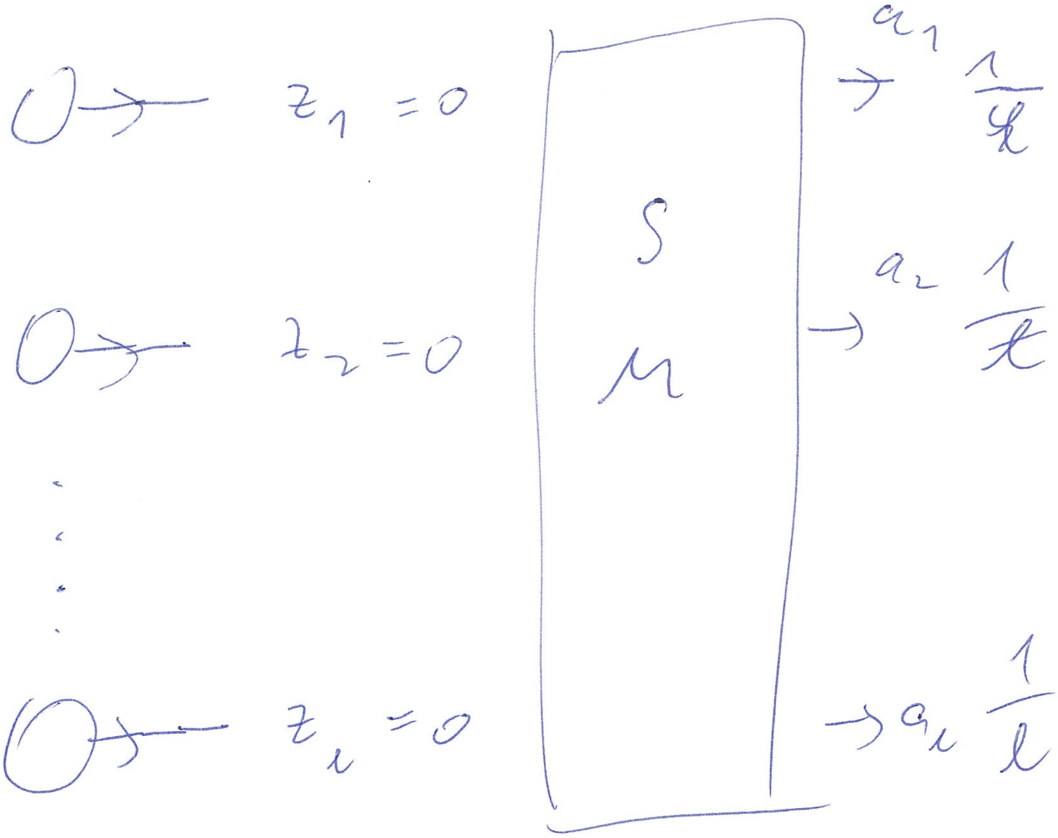
uncertainty of prob. distr.

Softmax
normalisation over multiple-classes

$$\sigma_{SM}(z_j) = \frac{e^{z_j}}{\sum_l e^{z_l}}$$



a_j ~ probabilities



$$\mathcal{L}_{x\text{-}e}(\tilde{y}, y) = -\sum_l y_l \log_2(\tilde{y}_l)$$

$$\tilde{y} = \begin{bmatrix} 0.1 \\ 0.173 \\ 0.56 \\ 0.167 \end{bmatrix}, \quad y = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\mathcal{L}_{x\text{-}e} = -0 \times \log_2(0.1) - 1 \cdot \log_2(0.173) + $$
$$- 0 \times \log_2(0.56) - 0 \cdot \log_2(0.167)$$

$$= -\log_2(0.173) \approx 2.53$$

$$\mathcal{L}_{MSE} = 0.68$$

3□ 1△

2□ 1△

Gini impurity

$G_1 = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2$ | $G_2 = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2$

0.38

$= 0.45$

3□ · 3△

6□ 0△

$$G_3 = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 =$$

$$= 1 - \frac{1}{2} = 0.5$$

$$G_4 = 1 - 1 = 0$$

$$G = 1 - \sum_i p_i^2$$